*International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*
**Web Site: www.ijettcs.org Email: editor@ijettcs.org**
**Volume 6, Issue 1, January - February 2017** **ISSN 2278-6856**

# DATA ANALYTICS APPLICATION USED IN THE FIELD OF BIG DATA FOR SECURITY INTELLIGENCE

**Prof. Amar Nath Singh[1], Er. Anurag Pattanayak[2], Er. Gyanachanda Samantaray[2]**

[1,2]Gandhi Engineering College, Bhubaneswar, Odisha

## ABSTRACT

***The term Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety). Human beings now create 2.5 quintillion bytes of data per day. The rate of data creation has increased so much that 90% of the data in the world today has been created in the last two years alone.2 This acceleration in the production of information has created a need for new technologies to analyze massive data sets. The urgency for collaborative research on Big Data topics is underscored by the U.S. federal government's recent $200 million funding initiative to support Big Data research. This paper describes how the incorporation of Big Data is changing security analytics by providing new tools and opportunities for leveraging large quantities of structured and unstructured data.***

**Key Words:-**Data Analytics, Big data, Traditional approach, Collaborative approach, Structured and Unstructured data.

## 1. INTRODUCTION:

As we know that, the term Big data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies [2]. In the current days application we are generating the huge amount of heterogeneous complex data which need a lot of strategy for its processing. In traditional days the appearance of data was small, which require less processing speed due to its arrangement. But now a day we need a high speed processing environment.

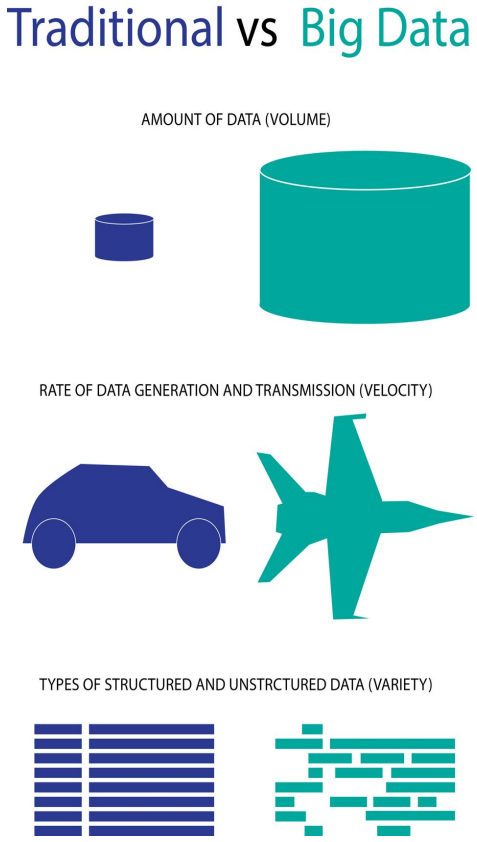


**Figure 1.** Big Data differentiators with traditional data.

Hence we need a proper concrete solution for processing of such data. So we have to go for the analysis of big data first before the processing.

## 2. BIG DATA ANALYTICS

The analysis of big data is very essential aspect now a day. Big Data analytics can produce operational and business knowledge at an unprecedented scale and specificity[4]. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools.

The technological advances in storage, processing, and analysis of Big Data include (a) The rapidly decreasing cost of storage and CPU power in recent years [1];

(b) The flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage [3]; and
(c) The development of new frameworks such as Hadoop, which allow users to take advantage of these, distributed computing systems storing large quantities of data through flexible parallel processing [4].
Hence, by using this approach, the traditional approach is now a day's no longer used.

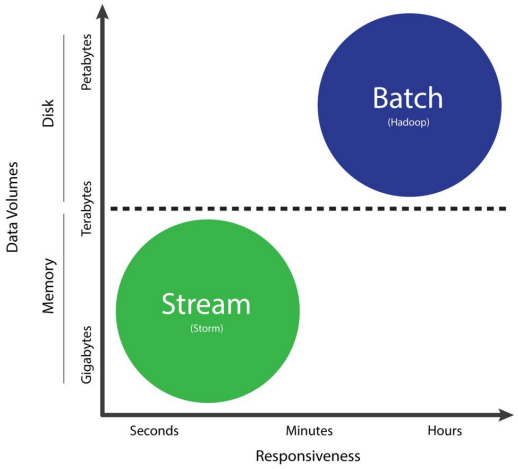


**Figure 3.** Batch and stream processing

Hadoop is one of the most popular technologies for batch processing. The Hadoop framework provides developers with the Hadoop Distributed File System for storing large files (Figure 3), which is tailored for frequently occurring large-scale data processing problems that can be distributed and parallelized.

### 2.1 Data Privacy and Governance

The preservation of privacy largely relies on technological limitations on the ability to extract, analyze, and correlate potentially sensitive data sets [5]. However, advances in Big Data analytics provide tools to extract and utilize this data, making violations of privacy easier. As a result, along with developing Big Data tools, it is necessary to create safeguards to prevent abuse [7].

In addition to privacy, data used for analytics may include regulated information or intellectual property. System architects must ensure that the data is protected and used only according to regulations.

The scope of Big Data can improve information security best practices. CSA is committed to also identifying the best practices in Big Data privacy and increasing awareness of the threat to private information [8]. CSA has specific working groups on Big Data privacy and Data Governance, and we will be producing white papers in these areas with a more detailed analysis of privacy issues.

## 3.Big Data Analytics for Security

This section explains how Big Data is changing the analytics landscape. In particular, Big Data analytics can be leveraged to improve information security and situational awareness. For example, Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view.

Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for Big Data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses.

In the context of data analytics for intrusion detection, the following evolution is anticipated:

- 1st generation: Intrusion detection systems – Security architects realized the need for layered security (e.g., reactive security and breach response) because a system with 100% protective security is impossible.
- 2nd generation: Security information and event management (SIEM) – Managing alerts from different intrusion detection sensors and rules was a big challenge in enterprise settings. SIEM systems aggregate and filter alarms from many sources and present actionable information to security analysts.
- 3rd generation: Big Data analytics in security (2nd generation SIEM) – Big Data tools have the potential to provide a significant advance in actionable security intelligence by reducing the time for correlating, consolidating, and contextualizing diverse security event information, and also for correlating long-term historical data for forensic purposes.

## 4.Big Data analytics used for security purposes

### 4.1Network Security

In a recently published case study, Zions Bancorporation[8] announced that it is using Hadoop clusters and business intelligence tools to parse more data more quickly than with traditional SIEM tools [6]. In their experience, the quantity of data and the frequency analysis of events are too much for traditional SIEMs to handle alone. In their traditional systems, searching among a month's load of data could take between 20 minutes and an hour. In their new Hadoop system running queries with Hive, they get the same results in about one minute.

### 4.2 Enterprise Events Analytics

Enterprises routinely collect terabytes of security relevant data (e.g., network events, software application events, and people action events) for several reasons, including the need for regulatory compliance and post-hoc forensic analysis [4]. Unfortunately, this volume of data quickly becomes overwhelming. Enterprises can barely store the data, much less do anything useful with it. For example, it is estimated that an enterprise as large as HP currently (in 2013) generates 1 trillion events per day, or roughly 12 million events per second. These numbers will grow as enterprises enable event logging in more sources, hire more employees, deploy more devices, and run more software.

### 4.3 Advanced Persistent Threats Detection

An Advanced Persistent Threat (APT) is a targeted attack against a high-value asset or a physical system. In contrast to mass-spreading malware, such as worms, viruses, and Trojans, APT attackers operate in "low-and-slow" mode. "Low mode" maintains a low profile in the networks and "slow mode" allows for long execution time [5]. APT attackers often leverage stolen user credentials or zero-day exploits to avoid triggering alerts. As such, this type of attack can take place over an extended period of time while the victim organization remains oblivious to the intrusion.

## 5.Data Sharing and Provenance

Experimental research in cyber security is rarely reproducible because today's data sets are not widely available to the research community and are often insufficient for answering many open questions. Due to scientific, ethical, and legal barriers to publicly disseminating security data, the data sets used for validating cyber security research are often mentioned in a single publication and then forgotten. The "data wishlist" (Camp, 2009) published by the security research community in 2009 emphasizes the need to obtain data for research purposes on an ongoing basis [4]. These data sets include anti-virus telemetry and intrusion-protection telemetry, which record occurrences of known host-based threats and network-based threats, respectively. The binary reputation data set provides information on unknown binaries that are downloaded by users who participate in Download Insight, Symantec's reputation-based security program. The history of binary reputation submissions can reveal when a particular threat has first appeared and how long it existed before it was detected. Similarly, the binary stability data set is collected from the users who participate in the Performance Insight program, which reports the health and stability of applications before users download them. This telemetry data set reports application and system crashes, as well as system lifecycle events.

## 6.Conclusions

The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential. The following are only some of the questions that need to be addressed:

**1. Data provenance:** authenticity and integrity of data used for analytics. As Big Data expands the sources of data it can use, the trustworthiness of each data source needs to be verified and the inclusion of ideas such as adversarial machine learning must be explored in order to identify maliciously inserted data.

**2. Privacy:** we need regulatory incentives and technical mechanisms to minimize the amount of inferences that Big Data users can make. CSA has a group dedicated to privacy in Big Data and has liaisons with NIST's Big Data working group on security and privacy. We plan to produce new guidelines and white papers exploring the technical means and the best principles for minimizing privacy invasions arising from Big Data analytics.

**3. Securing Big Data stores:** this document focused on using Big Data for security, but the other side of the coin is the security of Big Data. CSA has produced documents on security in Cloud Computing and also has working groups focusing on identifying the best practices for securing Big Data.

**4. Human-computer interaction:** Big Data might facilitate the analysis of diverse sources of data, but a human analyst still has to interpret any result. Compared to the technical mechanisms developed for efficient computation and storage, the human-computer interaction with Big Data has received less attention and this is an area that needs to grow. A good first step in this direction is the use of visualization tools to help analysts understand the data of their systems.

## Author

Prof. Amar Nath Singh is presently working at GEC, Bhubaneswar as a reader in the department of Computer Science and Engineering. His research area is Underground Mines and Surface mining using Artificial Intellegence, Fuzzy Logic. His research area includes cloud computing, WSN, Machine Learning. He has produced more than 60 M.Tech scholar till date.

## References

[1]. Alperovitch, D. (2011). Revealed: Operation Shady RAT. Santa Clara, CA: McAfee.

[2]. Bilge, L. & T. Dumitras. (2012, October) Before We Knew It: An empirical study of zero-day attacks in the real world. Paper presented at the ACM Conference on Computer and Communications Security (CCS), Raleigh, NC.

[3]. Bryant, R., R. Katz & E. Lazowska. (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science and society. Washington, DC: Computing Community Consortium.
[4]. Camp, J. (2009). Data for Cybersecurity Research: Process and "whish list". Retrieved July 15, 2013, from http://www.gtisc.gatech.edu/files_nsf10/data-wishlist.pdf.
[5]. Cugoala, G. & Margara, A. (2012). Processing Flows of Information: From Data Stream to Complex Event Processing. ACM Computing Surveys 44, no. 3:15.
[6]. Curry, S. et al. (2011). RSA Security Brief: Mobilizing intelligent security operations for Advanced Persistent Threats. Retrieved July 15, 2013, from http://www.rsa.com/innovation/docs/11313_APT_BRF_0211.pdf
[7]. Dumitras, T. & P. Efsathopoulos. (2012, May). The Provenance of WINE. Paper presented at the European Dependable Computing Conference (EDCC), Sibiu, Romania.
[8]. Dumitras, T. & D. Shou. (2011, April). Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Enviornment (WINE). Paper presented at the EuroSys BADGERS Workshop, Salzburg, Austria.